D230–D232 Nucleic Acids Research, 2005, Vol. 33, Database issue
doi:10.1093/nar/gki029

REBASE—restriction enzymes and DNA methyltransferases

Richard J. Roberts*, Tamas Vincze, Janos Posfai and Dana Macelis

New England Biolabs, Inc., 32 Tozer Road, Beverly, MA 01915 USA

Received September 17, 2004; Accepted September 22, 2004

ABSTRACT

REBASE is a comprehensive database of information about restriction enzymes, DNA methyltransferases and related proteins involved in restriction–modification. It contains both published and unpublished work with information about recognition and cleavage sites, isoschizomers, commercial availability, crystal and sequence data. Experimentally characterized homing endonucleases are also included. Additionally, REBASE contains complete and up-to-date information about the methylation sensitivity of restriction endonucleases. An extensive analysis is included of the restriction–modification systems that are predicted to be present in the sequenced bacterial and archaeal genomes from GenBank. The contents of REBASE are available by browsing from the web (http://rebase.neb.com/rebase/rebase.html) and through selected compilations by ftp (ftp.neb.com) and as monthly updates that can be requested via email.

INTRODUCTION

Since the last description of REBASE in the 2003 NAR Database Issue (1), there has been considerable growth in the size of the database primarily due to the large number of restriction–modification (RM) genes that can be found in the sequence databases. More than 200 bacterial and archaeal genomes are available from GenBank (2) and it is now clear that RM systems are much more common than had once seemed likely. Mainly, this is because of the difficulty of detecting Type I systems or solitary DNA methyltransferases by biochemical or genetic assay. Putative RM genes identified in these genomes are named systematically according to recently published nomenclature rules (3) and all have the suffix 'P' to indicate their putative status. The REBASE website (http://rebase.neb.com/rebase/rebase.html) summarizes all information known about every restriction enzyme and their associated proteins. This includes source, commercial availability, sequence data, crystal structure information, cleavage sites, recognition sequences, isoschizomers and methylation sensitivity. Within the reference section of REBASE, links are maintained to the full text of all papers whenever that is freely available on the web. Also, there is an extensive reciprocal cross-referencing between REBASE and NCBI. REBASE includes links to GenBank and PubMed, and NCBI's Linkout utility uses REBASE, PubMed and GenBank record numbers to hook directly into REBASE's enzyme, sequence, reference and genome data. Links to other major databases such as SwissProt (4), PDB (5) and PFam (6) are also maintained.

There are currently 3681 biochemically characterized restriction enzymes in REBASE and of the 3612 Type II restriction enzymes, 588 are commercially available, including 221 distinct specificities from a total of 253 total specificities known. As can be seen from Figure 1, the major growth in REBASE during the previous two years has been in the number of putative genes for RM system components. More than 620 restriction enzyme genes and 2200 DNA methyltransferase genes can be identified in GenBank entries. The sequenced microbial genomes provide more than 1700 of these genes.

The method used to identify putative RM genes in DNA sequences has three fundamental components: the REBASE database itself, an expert-derived set of RM system features and a computer program designed to spot these features in anonymous sequences. Each sequence analyzed is checked for its overall sequence similarity to REBASE gene sequences. For DNA methyltransferase sequences, which are the primary indicator of an RM system, the presence, proper order and characteristic spacing of well-conserved motifs suggest the candidates. The more widely divergent genes of the restriction enzymes reside close to the genes for their cognate methyltransferases. Such associations point to potential restriction enzyme genes, even when they lack any similarity to genes of known enzymes. Publicly available sources of non-eukaryotic sequences are also analyzed frequently by this system. All genes are manually inspected by a curator before entry into REBASE.

REBASE has its own dedicated webserver and can be searched extensively. Specialized information is available from the REBASE Lists icon and information about the

*To whom correspondence should be addressed. Tel: +1 978 927 3382; Fax: +1 978 921 1527; Email: roberts@neb.com

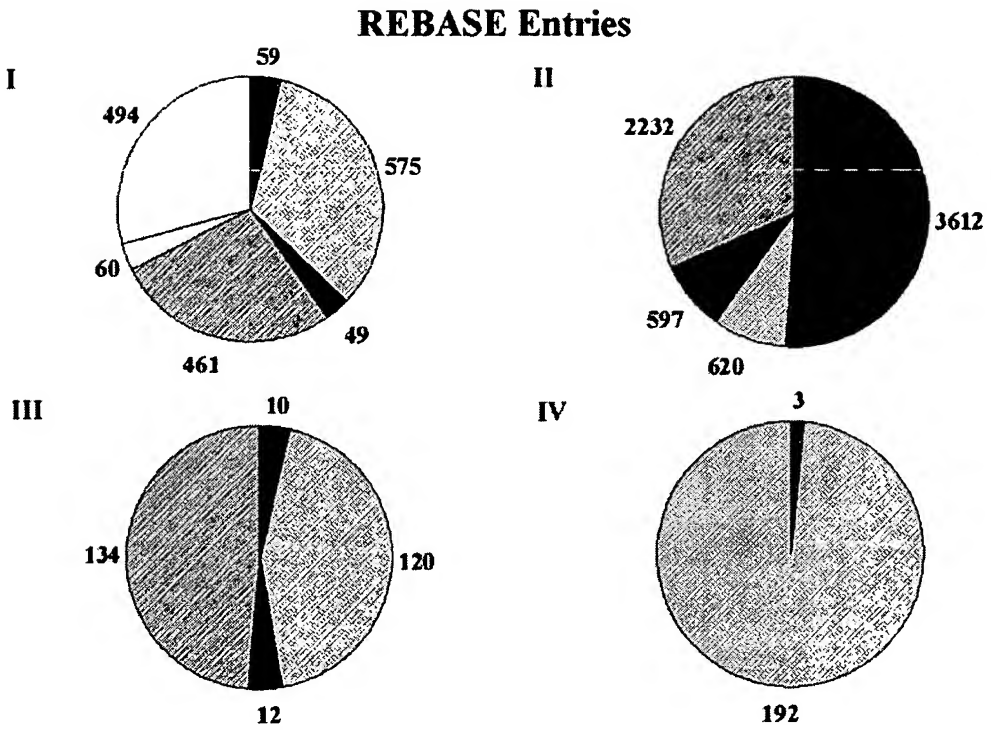

Figure 1. The circles depict the four major types of restriction enzymes and show the occurrence of their components in REBASE. The Type I enzymes have three genes encoding a methyltransferase subunit, a restriction subunit and a specificity subunit. The Type II enzymes have two separate genes, one encoding a restriction enzyme and one encoding a methyltransferase. The Type III enzymes also have two genes, one encoding a methyltransferase, which can operate alone or form part of a complex with a restriction subunit. The Type IV enzymes act alone and restrict DNA that is methylated. The solid colors indicate REBASE entries that have been experimentally verified and the shaded colors indicate genes predicted computationally. R genes are in red, M genes in blue and specificity genes in yellow. The numbers refer to the contents of REBASE on September 15, 2004.

sensitivity of restriction enzymes to DNA methylation can be found by clicking on the REBASE Methylation Sensitivity Icon. In the latter case, the data is shown in double-strand format so that the effects of hemi-methylation and double-strand methylation are clearly differentiated. REBASE also has links to useful programs via the REBASE Tools icon. NEBcutter analyzes DNA sequences for the presence of restriction enzyme recognition sites (7). REBsites will generate theoretical digests of an input DNA with each of the 253 known specificities. REBpredictor is a tool for predicting restriction enzyme recognition sites that is an updated version of TABLES (8) and a specific BLAST (9) option permits a new sequence to be analyzed for RM genes.

The REBASE Genomes Icon leads to data for the currently sequenced 193 bacterial and 21 archaeal genomes. Schematic representations of the whole genomes and the individual RM system within them are available and, from the pages showing the sequence schematics, there are links to the major database entries for these genes as well as links that will identify the closest neighboring sequences. This can be extremely useful in making predictions about the recognition sequence specificity of newly sequenced systems. This whole section of REBASE provides a valuable resource for the annotation of the RM genes in a newly sequenced bacterial genome, particularly given the large numbers of RM systems that are often found. Scientists interested in using the sequence information in REBASE to annotate microbial genomes are encouraged to contact the REBASE staff.

ACKNOWLEDGEMENTS

Special thanks are due to the many individuals who have so kindly contributed their unpublished results for inclusion in this compilation and to the REBASE users who continue to steer our efforts with their helpful comments. We are especially grateful to Karen Otto for secretarial help. This database is supported by the National Library of Medicine (LM04971) and New England Biolabs.

REFERENCES

1. Roberts,R.J., Vincze,T., Posfai,J. and Macelis,D. (2003) REBASE restriction enzymes and methyltransferases. *Nucleic Acids Res.*, **31**, 418–420.
2. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2004) GenBank: update. *Nucleic Acids Res.*, **32**, D23–D26.
3. Roberts,R.J., Belfort,M., Bestor,T., Bhagwat,A.S., Bickle,T.A., Bitinaite,J., Blumenthal,R.M., Degtyarev,S.Kh., Dryden,D.T.F., Dybvig,K. *et al.* (2003) A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res.*, **31**, 1805–1812.
4. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.*

(2003) The SWISS-PROT protein knowledge base and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
5. Bourne,P.E., Addess,K.J., Bluhm,W.F., Chen,L., Deshpande,N., Feng,Z., Fleri,W., Green,R., Merino-Ott,J.C., Townsend-Merino,W. *et al.* (2004) The distribution and query systems of the RCSB Protein Data Bank. *Nucleic Acids Res.*, **32**, D223–D225.
6. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonhammer,E.L.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
7. Vincze,T., Posfai,J. and Roberts,R.J. (2003) NEBcutter: a program to cleave DNA with restriction enzymes. *Nucleic Acids Res.*, **31**, 3688–3691.
8. Gingeras,T.R., Milazzo,J.P. and Roberts,R.J. (1978) A computer assisted method for the determination of restriction enzyme recognition sites. *Nucleic Acids Res.*, **5**, 4105–4127.
9. McGinnis,S. and Madden,T.L. (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.*, **31**, W20–W25.